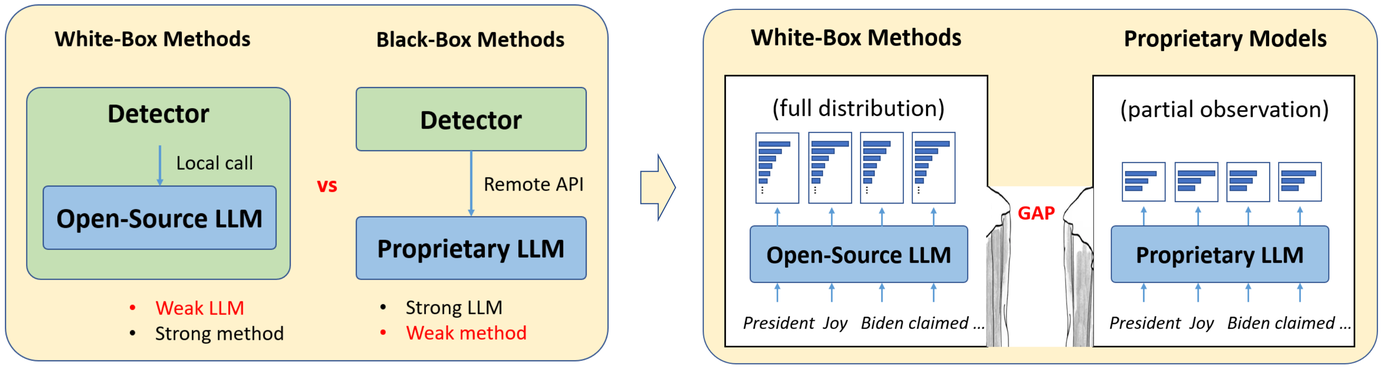
Glimpse: Enabling White-Box Methods to Use Proprietary Models for Zero-Shot LLM-Generated Text Detection

Guangsheng Bao, Yanbin Zhao, Juncai He, Yue Zhang

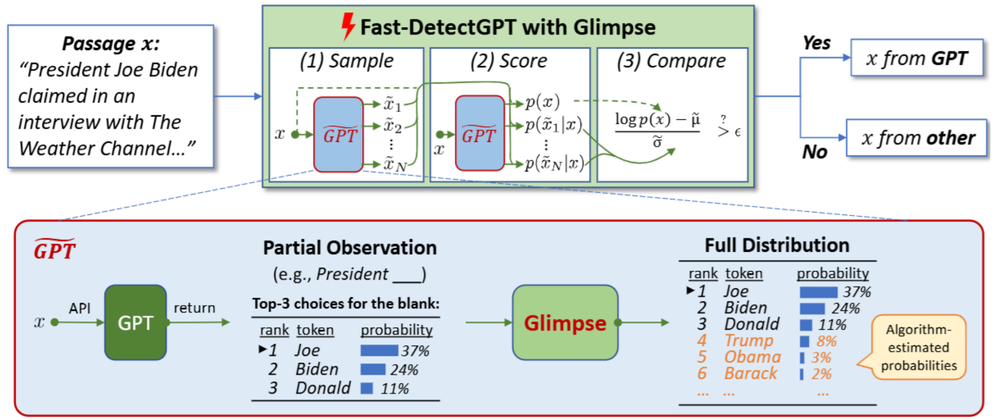
(baoguangsheng@westlake.edu.cn)

I. Motivation & Challenge



II. Method: Bridge the Gap

- **Glimpse**



- **Probability Distribution Estimation**

Geometric Distribution

$$\begin{cases} p(k) = p_k, & \text{for } k \in [1..K] \\ p(k) = p_K \cdot \lambda^{k-K}, & \text{for } k \in [K+1..M] \\ \sum_{k=1}^{M} p(k) = 1, \end{cases}$$

Zipfian Distribution

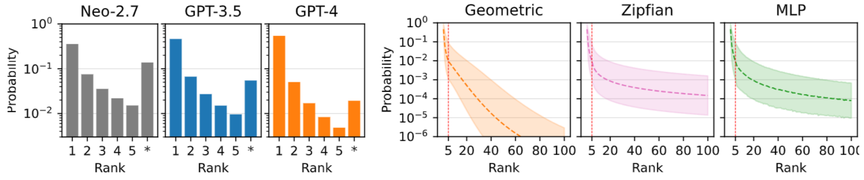
$$\begin{cases} p(k) = p_k, & \text{for } k \in [1..K] \\ p(k) = p_K \cdot (\frac{\beta}{k-K+\beta})^{\alpha}, & \text{for } k \in [K+1..M] \\ \sum_{k=1}^{M} p(k) = 1, \end{cases}$$

MLP Model

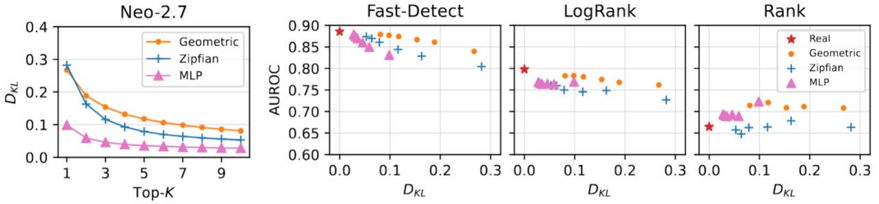
$$\begin{cases} p(k) = p_k, & \text{for } k \in [1..K] \\ p(k) = p_{\text{rest}} \cdot p_{\text{MLP}_\theta}(k-K), & \text{for } k \in [K+1..M] \\ \sum_{k=1}^{M} p(k) = 1, \end{cases}$$

III. Results & Analysis

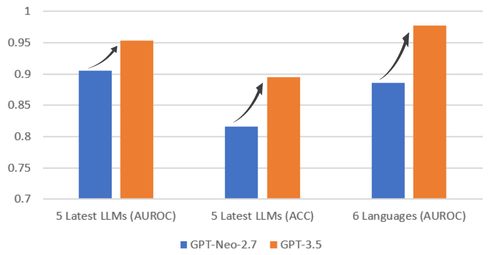
- **Analysis of Distribution**



- **Analysis of Effectiveness**



- **Main Results**



- **Conclusion**

➢ **Estimated distributions works at the same level as the real distributions.**

➢ **Strong proprietary LLMs are also strong detectors.**

Zhejiang University
Westlake University
Shanghai Polytechnic University
King Abdullah University of Science and Technology